Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 16:07:25
REFERENCE NO: 278

Author Names & Affiliations

- kc claffy - ucsd, sdsc, cada
- amogh dhamdhere - ucsd, sdsc, caida

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

computer science, Internet research, internet measurement

Title of Submission

Platform for Integrative Analysis and Visualization of Large-Scale Internet Measurement Data

Abstract (maximum ~200 words).

As the Internet and our dependence on it have grown, the structure and dynamics of the network, and how it relates to the political economy in which it is embedded, have gathered increasing attention by researchers, operators and policy makers. All of these stakeholders bring questions that they lack the capability to answer themselves. Epistemological challenges lie in developing and deploying measurement instrumentation and protocols, expertise required to soundly interpret and use complex data, lack of tools to synthesize different sources of data to reveal insights, data management cost and complexity, and privacy issues. Although a few interdisciplinary projects have succeeded, the current mode of collaboration simply does not scale to the exploding interest in scientific study of the Internet, nor to complex and visionary scientific uses of CAIDA's data by non-networking experts. We believe the community needs a new shared cyberinfrastructure resource that integrates active Internet measurement capabilities, multi-terabyte data archives, live data streams, heavily curated topology data sets revealing coverage and business relationships, and traffic measurements. Such a resource would enable a broad set of researchers to pursue new scientific directions, experiments, and data products that promote valid interpretations of data and derived inferences.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

For two decades, UC San Diego's Center for Applied Internet Data Analysis (CAIDA) has been developing data-focused products, services, open source software tools, and resources to advance the field of Internet science. This field has permeated disciplines ranging from theoretical computer science to political science, from physics to techlaw, and from network architecture to public policy. As the Internet and our dependence on it have grown, the structure and dynamics of the network, and how it relates to the political economy in which it is embedded, has gathered increasing attention by researchers, operators and policy makers. All of these stakeholders bring questions that they lack the capability to answer themselves. Epistemological challenges lie in developing and deploying measurement instrumentation and protocols, expertise required to soundly interpret and use complex data, lack of tools to synthesize different sources of data to reveal insights, data management cost and complexity, and privacy issues. As it has become clear how many fears and aspirations about the Internet (security, affordability, neutrality, universal service, congestion) are rooted in economics, ownership, and trust issues, CAIDA has cultivated communities of economists, law, and policy researchers with an interest in empirically grounding their understanding of the Internet. But the impact thus far has been limited to a handful of researchers. The current mode of collaboration simply does not scale to the exploding interest in scientific study of the Internet, nor to complex and visionary uses of CAIDA's data by non-networking experts in other fields of science and engineering.

In response to feedback from these communities, as well as our own insights from meetings, workshops, and other discussion forums, we believe the research community needs a new shared cyberinfrastructure resource - a Platform for Applied Network Data Analysis (PANDA). This system would leverage existing research infrastructure measurement and analysis components, that, once connected, will enable new scientific directions, experiments, and data products. The idea for this platform builds on input from dozens of researchers over seven years of CAIDA workshops, and we have developed an initial design plan that supports its use, evaluation, extensibility, and sustainability. The proposed system would integrate active Internet measurement capabilities, multi-terabyte data archives, live data streams, heavily curated data sets revealing coverage and business relationships, and traffic measurements that represent, for many researchers, the holy grail of scientific sources of information about Internet behavior. Such a resource would enable a broad set of researchers to access, query, visualize, and analyze Internet data, as well as create new data products in ways that promote valid interpretations of data and derived inferences. We also need visualization tools to allow non-experts to understand various aspects of Internet structure, using geographic and economic annotations on the data, with access controls where appropriate for sensitive data.

This idea draws inspiration in part from an NSF-funded Internet research (NeTS) project at CAIDA that would immediately benefit from such system integration: the Measurement and ANalysis of Internet Congestion [1]

(MANIC) project gathers time series of performance data for tens of thousands of interconnection links (connecting networks to each other), giving researchers and policymakers visibility into an ecosystem that has long been shrouded in secrecy. Demand for such quantitative insight is growing, but we have only scratched the surface of what the instrumentation we have built can reveal. To allow the MANIC project achieve its full potential contributions to both scientific research and empirical grounding of public policy, we must expand its scope to many more measurement vantage points, merge in external data sources, and provide various stakeholders the ability to analyze and visualize the data. Doing so involves solving significant cyberinfrastructure challenges in large data analytics and scalable techniques to intuitively visualize multi-resolution graph structures with complex annotations (e.g., time series, geolocation, and performance metrics) on nodes, edges, and sub-graphs. These research objectives are not unique to the MANIC project, and we envision a set of solutions that could extend to other Internet measurement problems in cybersecurity (e.g., detecting attacks or routing hijacks), network management and troubleshooting (e.g., detecting network outages and routing instabilities), and policy (e.g., measuring broadband access bandwidth, coverage, and user Quality of Experience).

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Understanding the Internet as critical infrastructure - properties of its structure, dynamics, performance, vulnerabilities - although still a young field of science, is maturing rapidly. However other fields of information technology and engineering are evolving even more rapidly, leveraging Moore's law and the open-software revolution to advance data processing, big data analytics, and visualization methods and networked system designs. Our fortunate position at the UC San Diego Supercomputer Center has given us visibility into both communities, and revealed a surprising gap that offers a compelling opportunity: integration of cutting-edge tools for data analytics, text processing and visualization, all backed by existing HPC resources into a platform that can transform Internet scientific research for a broad cross-section of the community.

Such a research-driven systems integration effort could amplify the return on NSF's ongoing commitment to R&E cyberinfrastructure investments, while supporting three ongoing technical challenges in Internet research. First, many Internet measurement projects produce large volumes of time series data which they need to efficiently analyze, correlate, and visualize via public facing query interfaces. Second, applications in network measurement, security, and network operations involve mining text-based data consisting of billions of records to track changes or detect anomalies. But the tools used to process text-based measurement data are fairly rudimentary, and do not take advantage of tremendous recent advances in the field of document search and indexing. Finally, an important class of Internet measurement data consists of dynamic, multi-resolution graphs with annotations on nodes, links and subgraphs along various dimensions (geography, performance, economics). The research community lacks a platform to support exploratory visualization of Internet-scale datasets.

The technical challenges include building data collection, processing, and analysis systems that combine heterogeneous data from diverse datasets to enable agile analytics on millions of time series data in streaming and batch modes. Additionally we need to apply recent advances in text search to mine insights from Internet measurement datasets that can be represented as text documents (routing, DNS, topology, traffic, security logs). Finally, we need new techniques to visualize large, multi-resolution, annotated, evolving graph structures, allowing many users to interactively browse the data in an intuitive manner. Ideally, within five years these data science technologies could be integrated into a unified platform for data ingestion, management, and analytics, that are designed to leverage current and emerging high performance computing, storage, and networking capabilities.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Other considerations in developing such cyberinfrastructure include sustainable collection, curation, and storage of large volumes of complex data, including data volume and quality, validation of data and inferences, mechanisms for enabling privacy-respecting data sharing, and training students and researchers to promote ethical use of data in research. These considerations are now pertinent to all fields of science and engineering. By creating a platform that is likely to inspire classroom and student project use, we will have a laboratory to facilitate testing data handling methods from other fields (e.g., biomedical research) as well as offering new methods for emerging fields to consider.

As NSF's previous investments have demonstrated, cyberinfrastructure designed and managed to maximize its reach and utility will have benefits that transcend the disciplines who initially use it. Exemplifying this tradition, a platform for scientific analysis of the evolving Internet infrastructure will not only advance Internet science today, but help prepare tomorrow's workforce to better understand and navigate the network infrastructure that, despite all its risks, increasingly mediates our lives.

Consent Statement